How to train RNNs on chaotic neural data

Jonas Mikhaeil*, Zahra Monfared*, Daniel Durstewitz

Department of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University

Introduction

RNNs are a promising and increasingly popular tool for the analysis of ***neural recordings*** [2,3]

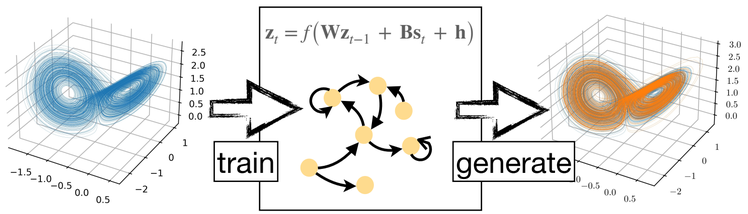
To trust in and interpret results obtained by training RNNs on neural data, we need to be certain they recover the underlying dynamics
-> ***dynamical systems reconstruction***

Training RNNs is notoriously hard, especially if the data are governed by long and complicated temporal dependencies
-> ***exploding and vanishing gradient problem*** [1,4]

Problem setting

Reconstruction of **chaotic** dynamical systems (DS)

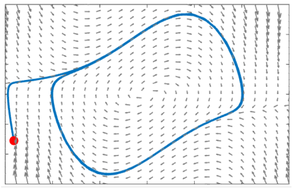
Observed DS | RNN | Reconstructed DS



Training RNNs on chaotic data is **hard**

chaotic dynamics → **exploding gradients during training**

Equilibrium or cyclic behaviour
lead to vanishing gradients



Theorem: *Consider a generic RNN and assume that it converges to a stable FP or k-cylce Γ_k ($k \geq 1$) with $\mathscr{B}_{\Gamma_k}$ as its basin of attraction. Then for every $\mathbf{z}_1 \in \mathscr{B}_{\Gamma_k}$ the Jacobian $\frac{\mathbf{z}_T}{\partial \mathbf{z}_t}$ vanishes as $T \to \infty$. Moreover, the tangent vector $\frac{d\mathbf{z}_t}{dw_{mk}}$ and thus the gradient of the loss function $\frac{\partial \mathscr{L}_T}{\partial \mathbf{W}}$ will be bounded, i.e. will not diverge as $T \to \infty$.*

This work was funded by the German Science Foundation (DFG) through individual grant Du 354/10-1 to DD, and via the Excellence Cluster 'Structures' at Heidelberg University (EXC-2181 – 390900948)

Theoretical Framework

Jacobians required for training and RNN long-term behaviour are closely linked

Training

Jacobians $\frac{\partial \mathbf{z}_t}{\partial \mathbf{z}_k} = \prod_{r=0}^{T-k-1} \frac{\partial \mathbf{z}_{T-r}}{\partial \mathbf{z}_{T-r-1}} = \prod_{r=0}^{T-k-1} \mathbf{J}_{T-r}$

connecting temporally distal states

$\lambda_{\max} > 0$

chaotic behaviour
(for bounded orbits)

$\lambda_{\max} > 0 \Rightarrow \lim_{T\to\infty} \left\| \prod_{t=0}^{T-2} J_{T-t} \right\| = \lim_{T\to\infty} \left\| \frac{\partial \mathbf{z}_T}{\partial \mathbf{z}_t} \right\| = \infty$

causes ***exploding gradients***

Dynamics

Lyapunov exponent $\lambda_{\max} = \lim_{T\to\infty} \left(1/T \log \left\| \prod_{r=0}^{T} \mathbf{J}_{T-r} \right\|\right)$

determines long-term behaviour

$\lambda_{\max} < 0$

fixed point or cylce

$\lambda_{\max} < 0 \Rightarrow \lim_{T\to\infty} \left\| \prod_{t=0}^{T-2} J_{T-t} \right\| = \lim_{T\to\infty} \left\| \frac{\partial \mathbf{z}_T}{\partial \mathbf{z}_t} \right\| = 0$

creates ***vanishing gradients***

Training with exploding gradients:
Sparsely forced BPTT

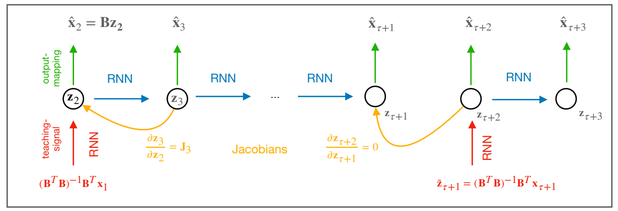
Force the RNN every τ time-steps onto observations

$$\mathbf{z}_{t+1} = \begin{cases} RNN(\tilde{\mathbf{z}}_t) \text{ if } t \in \mathscr{T} = \{n\tau + 1\}_{n\in\mathbb{N}_0} \\ RNN(\mathbf{z}_t) \text{ else} . \end{cases}$$

This training procedure
1) breaks trajectory divergence
2) alleviates the exploding gradient problem

The forcing signal is obtained by "inverting" the output layer

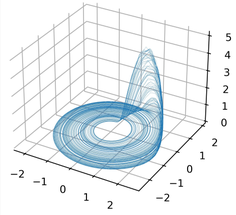
$\tilde{\mathbf{z}}_t = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x}_t$



The learning interval τ can be chosen in accordance with the Lyapunov spectrum

$\tau_{pred} = \log(2)/\lambda_{\max}$

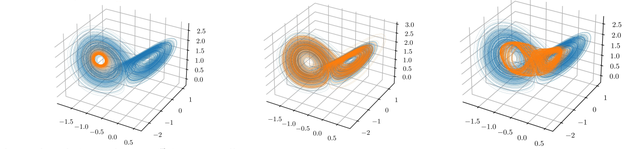
Chaotic dynamics
lead to exploding gradients



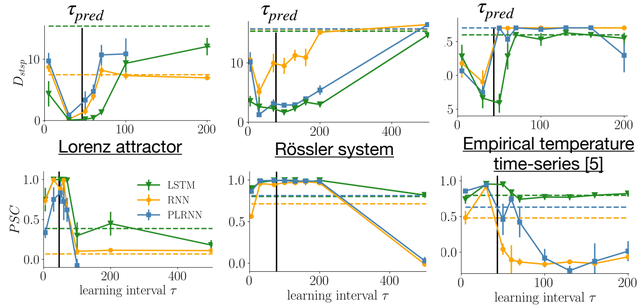
Theorem: *Suppose that a generic RNN has a chaotic attractor Γ^* with $\mathscr{B}_{\Gamma^*}$ as its basin of attraction. Then, for every orbit with $\mathbf{z}_1 \in \mathscr{B}_{\Gamma^*}$ the Jacobians $\frac{\partial \mathbf{z}_T}{\partial \mathbf{z}_t}$ connecting temporally distal states $\mathbf{z}_T$ and $\mathbf{z}_t$ ($T \gg t$), the tangent vectors $\frac{d\mathbf{z}_T}{dw_{mk}}$ and thus the gradients of the loss function $\frac{\partial \mathscr{L}_T}{\partial \mathbf{W}}$ will explode for $T \to \infty$.*

Results:

Reconstruction examples:
blue: data, orange: LSTM reconstruction



learning interval $\tau = 5$ too small | $\tau = 30$ optimal | $\tau = 200$ too large



(D_{stsp} : overlap in attractor geometry, PSC : power spectrum correlations)

Sparsely forced BPTT allows the reconstruction of chaotic dynamical systems
(continues lines)

and significantly outperforms the classical approach of gradient clipping
(dashed lines)

References

[1] Y. Bengio, P. Simard, and P. Frasconi, 'Learning long-term dependencies with gradient descent is difficult', IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, Mar. 1994, 10.1109/72.279181.
[2] C. Pandarinath et al., 'Inferring single-trial neural population dynamics using sequential auto-encoders', Nat Methods, vol. 15, no. 10, pp. 805–815, Oct. 2018, 10.1038/s41592-018-0109-9.
[3] M. G. Perich et al., 'Inferring brain-wide interactions using data-constrained recurrent neural network models', Neuroscience, preprint, Dec. 2020, 10.1101/2020.12.18.423348.
[4] D. Schmidt et al., 'Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies', ICLR, 2021, https://openreview.net/forum?id=_XYzwxPlQu6
[5] recorded at the Weather Station at the Max Planck Institute for Biogeochemistry in Jena, Germany spanning the time period between 2009 and 2016. Data can be accessed at https://www.kaggle.com/pankrzysiu/weather-archive-jena